DOCUMENT RESUME

ED 435 751 TM 030 403

AUTHOR	Le Rouzie, Violaine; Cusick, Mary
TITLE	Immediate Evaluation of Training Events at the Economic Development Institute of the World Bank Measuring Reaction, Self-Efficacy and Learning in a Worldwide Context.
PUB DATE	1999-11-00
NOTE	10p.; Paper presented at the Annual Meeting of the American Evaluation Association (Orlando, FL, November 3-6, 1999).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Evaluation Methods; *Learning; Program Evaluation; Responses; *Self Efficacy; Self Evaluation (Individuals); *Training
IDENTIFIERS	*World Bank

ABSTRACT

The Economic Development Institute of the World Bank (EDI) conducts extensive training events for a variety of audiences throughout the world. Over the years, EDI has used several types of instruments to evaluate these events. Drawing lessons from these experiences, this paper presents a "hierarchy" of training evaluation designs. These are: (1) one-time post-training design; (2) pre-post self-assessment; and (3) pre-post design for assessing learning. The pros and cons of self-assessment versus actual testing, one-time measure versus multiple measures, and how to introduce a "control group" in a feasible way for a training organization are discussed. The paper also addresses some measurement challenges due to the context of international development involving worldwide audiences. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 435 751

Immediate Evaluation of Training Events at the Economic Development Institute of the World Bank Measuring Reaction, Self-Efficacy and Learning in a Worldwide Context

Violaine Le Rouzie
Mary Cusick

TM030403

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
V. Le Rouzie
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Presented at the
American Evaluation Association Annual Conference 1998
Chicago, November 1998

BEST COPY AVAILABLE

Immediate Evaluation of Training Events at the Economic Development Institute of The World Bank Measuring Reaction, Self-Efficacy and Learning in a Worldwide Context

Authors: Violaine Le Rouzic and Mary Cusick
Evaluation Unit of the Economic Development Institute of the World Bank

Presented at the
American Evaluation Association Annual Conference 1998
in Chicago on November 7, 1998

Abstract

The Economic Development Institute of the World Bank (EDI), conducts extensive training events for a variety of audiences throughout the world. Over the years, EDI has used several types of instruments to evaluate these events. Drawing lessons from this experience, the paper will present a "hierarchy" of training evaluation designs. The pros and cons of self-assessment versus actual testing, one-time measure versus multiple measures, and how to introduce a "control group" in a feasible way for a training organization will be discussed. This paper will then address some measurement challenges due to a context of international development involving worldwide audiences.

Introduction

This presentation—based on some experiences of the Economic Development Institute of World Bank in evaluating its training events during the events themselves—is divided into two parts. First, we will introduce an array of measurement designs and discuss in each case what the design measures, and what its strengths and limitations are. The second part will focus on how we deal with the challenges of our international context.

Before starting, we would like to give you an idea of our clients because our measurement strategy is (of course) built around them. We deal with two types of clients: internal and external.

a. Our internal clients are the trainers who organize the training events. We work with them to develop our evaluation instruments. We are a young internal evaluation unit trying to promote an evaluation culture within the Institute. Due to this context, we focus more on the formative side of the evaluation rather than on accountability.

b. Our external clients are the participants in the training events. They are well-educated audiences from all over the world. Most of them come from developing countries and countries in transition from centrally planned economies to market economies. The bulk of our training aims to discuss policy issues and to provide knowledge on how to assess and implement various policies in a wide range of development areas.

I. Types of designs, what they measure, their strengths and limitations:

This paper will present various evaluation designs using questionnaires administered to a class during training events. Questionnaires are the instruments most often used, although not exclusively.

A. One-time post-training design:

The one-time post training design asks participants for their reaction at the end of an event. It is widely understood to be a weak design. It provides no baseline data. In addition, one-time post-training design does not enable to evaluate the extent to which the learning objectives of the training have been achieved. However, it can be useful to gather the participants' reactions to some design features of the training event. The design is appropriate when the issue is to know what type of pedagogical method participants appreciate more. What the participants evaluate is by the end of the training known to them, and it is fresh in their minds.

This design was helpful when, five years ago, the Economic Development Institute of the World Bank started to undertake a massive training effort directed at the Former Soviet Union and Eastern European countries. Our trainers were divided on the issue of whether or not participatory pedagogical methods were appropriate for participants from the Former Soviet Union. Some trainers argued that Russians had been educated through heavily content-loaded lectures and would feel very uncomfortable if asked to engage in participatory methods, like role playing, for example. Others felt that people from the Former Soviet Union would be appreciate participatory methods to the same extent that Western countries appreciate them. Reactions to evaluation questions on the participants' appreciation of the methods used, along with their overall reactions to the training events, clearly indicated that participatory methods could be successful in the Former Soviet Union.

Another value of the one-time post-training design is that it is easy to design and administer questions. Questions need only be administered once at the end of a training event. Generally we do ask questions on the training design one time during or at the end of a training event, but we supplement them with other questions, notably pre-post self-assessment questions.

B. Pre-post self-assessment.

We currently use a pre-post design to assess the participants' level of confidence in their ability to use the materials provided or their level of knowledge on the contents of the training.

There are several dimensions to be cautious about with pre-post self-assessment designs.

First, there is the issue of when to administer the questionnaires. Sandi Mann (1997) conducted research that touched on this area. In theory, you could administer a pre-training questionnaire at the beginning of the training event and a post-training questionnaire at the end of the training. However, such a sequence of administration leads to unreliable data collection for the following reasons.

At the beginning of a training event, the participants do not usually have a clear idea of how much they know on a topic. Some participants may feel knowledgeable about a topic before a training class, therefore give a high rating to their perceived level of knowledge about this topic on the pre-training questionnaire. During the training class, the participants could discover that the topic is much more complex than anticipated. Humbled by this discovery, some participants could rate their knowledge of the topic rather low by the end of the class. The results could show a decrease in ratings rather than an increase, even though their post-training level of topic knowledge was higher than their pre-training level. On the other hand, some participants may feel that they know little about a topic before a class and discover that they actually knew the concept taught, but under a different terminology. The post-training ratings may be much higher than the pre-training ratings. This would not be due to the training itself, but to a misunderstanding of the question before the training.

To avoid the problem of when to administer the questionnaire, we use a "then-post design" instead of a pre-post design. We ask the participants to assess their level of confidence or knowledge by using only one administration of a questionnaire at the end of the training. We ask the participants to rate not only the level that they feel they have at this point in time, but also to assess, retrospectively, the level that they had just before the training event.

By using a retrospective assessment, the participants have a common understanding of the terms and of the universe of knowledge available. This common baseline of reference created during the training provides more reliable data. It also reduces the time spent in questionnaire administration. Finally, a then-post design makes it simpler to link the pre-training results to the post-training results for each participant because both sets of responses are on the same form.

The second aspect to consider when using a self-assessment design is that, as its name implies, it is only an assessment of one's own perceived level. Dixon (1990) showed that there is no significant correlation between what people feel they learned and what they actually learned. This research has been confirmed in our own observations of the phenomenon. We found that participants can either overestimate their knowledge or underestimate it.

Therefore we definitely do not recommend using this design to measure learning. However, considering our internal context (in which we have had to build a culture of evaluation), we sometimes had to use this design because the trainers initially refused to have their participants tested for their actual knowledge. Using this design proved to be disappointing to the trainers who chose it hoping to measure learning. As a result, when trainers repeat the training they often ask to use measures of actual learning. So as an intermediary step in a context of building an evaluation culture, this design can be useful.

We feel comfortable about using a self-assessed then-post design to measure the participants' level of confidence in their ability to use what they were taught at training. Confidence or self-efficacy is a more subjective dimension than learning. Self-assessment design is more useful in gauging self-efficacy than in measuring learning. In addition, self-efficacy is a valuable dimension to measure because according to Applebaum (1996) self-efficacy is a reliable predictor of motivation as well as skill performance.

The third element of caution to consider when using self-assessed design is that it is not only a self-assessment design, but also a self-reported assessment design. To make things even worse, the reporting is not spontaneous, but elicited by our questionnaires. We all know that some factors, such as social desirability or feelings towards the trainers, can, among others, influence the participants' ratings.

We try to account for these biases by introducing a question that elicits a then-post self-assessment on a topic that is not addressed in the training. We call it the "control question." We compare responses to the control question with responses to the list of questions on topics actually addressed by the training. The training-related questions are considered to be the "treatment questions." We control the potential biases due to self-reporting by simply following two steps: First we calculate the increase between the post-ratings and the then-ratings both for the treatment questions and for the control question. Then we deduct the increase obtained with the control question from the increase obtained with the treatment questions.

Although preliminary, our findings regarding the use of a control question are so far encouraging. They show an increase, but one smaller than that of the treatment questions. All of this is consistent with our expectations.

For a control question to be effective, it is important that it blends in with the list of treatment questions. It should not be obvious to the participants that the control question is actually assessing their rating behavior. For this, we usually advise the trainers to include a question on a topic that they initially intended to teach, but eventually had to cancel because of lack of time, or financial or human resources.

As we saw, self-assessment (with some caveats) can be a useful measure, but it should not replace the assessment of actual learning.

C. Pre-post design for assessing learning

To assess how much the participants learned in the training, we use a true pre-post design with two administrations of questionnaires: one at the beginning of the training and one at the end.

Contrary to the tests used in universities, learning assessments do not aim to measure the level reached by a particular individual. Due to the high level of our participants, most trainers are reluctant to the idea of testing individual's knowledge. They feel that it would be inappropriate to test senior level officials like students. We therefore target our evaluations to determine how much learning took place among the group of participants as a whole to see if the teaching was effective. Consequently, our learning assessments are anonymous.

However, we still need to match the responses to the pre-training assessment with those of the post-training assessment to ensure that the populations in both administrations of the test are identical. For this, we use an evaluation code number randomly assigned to individual participants. The participants have to indicate their individual evaluation code number on every evaluation questionnaire. This allows us to correlate the results of questions that were asked on different questionnaires, while protecting the anonymity of the respondents.

To ensure content validity, the trainers usually design forty multiple choice questions on the topics that they will teach. We review these questions and select the best thirty of them to be part of the learning assessment.

We then randomly assign these thirty questions to two groups: one will becomes the pre-test and the other one the post-test. The assignment of the questions into two different tests allows to avoid the pre-test bias that would occur if the same questions were asked before and after the training. In addition, the fact that the assignment is random ensures that the level of difficulty of both tests is even.

We can score the pre-test and the post-test and see if their difference is significant. Because randomization allowed to even out their level of difficulty, and because our training settings are usually intense and short, other threats to validity such as history are not a factor. Therefore we can conclude with confidence whether or not learning occurred due to the training.

The main limitation of this design is that we rely on multiple choice questions. This is the best way to achieve a reliable process for test scoring, but it also limits the cognitive domain tested. We reach the level of knowledge and comprehension of the participants, as defined by Bloom's Taxonomy of educational objectives, but not the level of analysis.

II. Challenges of the international context:

Let's look now at what constitutes the biggest challenge of the context in which we work: the fact that the respondents to our questionnaires come from all over the world.

A. Language used in the questions

Our international audience requires that we pay particular attention to the language used in our questionnaires. Many of our activities attract people from all over the world for training conducted in English. Although these people are usually highly educated, English is often their second, third or fourth language. They do not always master it perfectly. This requires us to use a language that can be understood by the largest number of people. I call this language the "modern Esperanto." It is basically English, but uses simple common words and words with Latin roots (to the extent possible). Latin words translated to English tend to be more elaborate than colloquial English. While the native English speakers are comfortable with this elaborate level of language, participants from developing countries who, besides English, use mostly French, Spanish or Portuguese as a national language (with Latin roots) find it easier to understand questions when simple English words with Latin roots are used.

An example of this is the wording of our scale of values for the numerical questions of our Level 1 evaluation. We chose "minimum" and "maximum" for the one-direction scale and "insufficient, adequate and excessive" for the items that deal with quantity and can have two negative poles. These words are easy to understand for an international audience and relatively easy to translate in various languages.

B. Clear instructions and presentations

Although our participants are highly educated people, they are often not accustomed to being surveyed. Answering questionnaires is a practice much more common in Western countries than in the rest of the world. Because of this, we need to pay particular attention to instructions and presentations.

To make our questionnaires look shorter, we used to have a stem question such as: "To what extent did the training help you to:" followed by several question endings. We expected people to rate each aspect by using the scale at the end of each question ending. However, we found that we were losing data because too many participants answered this type of question as a multiple choice question. Instead of choosing an answer at the end of each line they would select the letter at the beginning of the line. To avoid this problem, we now write each question to be rated in full.

We also have to be very clear in our instructions on how to indicate an answer. Different regions use different practices to select answers when given a scale. For example, not all countries use check marks to indicate an answer choice. Some countries use crosses. However, some countries use a cross to indicate the answer and other countries use a cross to eliminate the option that is not chosen. If you do not indicate what you want, you may end up with a cross over a YES and nothing over a NO and you will not be able to interpret the answer. If the questionnaires are not scannable, we ask people to circle their answer. If it is scannable they fill a circle.

C. Interpreting the results

Once we get the quantitative results, we also need to be aware that cultures do not all have the same pattern of ratings. We looked at the results of over 12,000 respondents from six different broad regions in the world: Sub Saharan Africa, East Asia and Pacific, Eastern Europe and Central Asia, Latin America and the Caribbean, Middle East and North Africa, and South Asia.

We found, based on our observations in regional and national seminars, that people from Middle East and North Africa and people from South Asia rate on average significantly lower than the rest of the world. This average is pulled down by the fact the a significant amount of them do not use the maximum possible ratings on the scale.

We could not explain these results by factors such as topics taught, the quality of the trainers or the level of education of the participants (which happened to be equivalent in all regions). We think that this might be due to cultural differences in rating pattern, although it would require further investigation to confirm it and eventually determine why.

Conclusion

Each design presented is appropriate for certain situations. One-time post-training design can measure participants' reaction to the pedagogy used. Then-post self-assessment completed by a control question can measure self-efficacy gains. Pre-post learning assessment measure actual learning. It is important to know what each design measures and, even more, what it does not measure in order to avoid drawing false conclusions. One characteristic of all measures discussed is that they are immediate and do not indicate what the mid- or long-term effect of the training will be.

Bibliography

Applebaum, Steven H. "Self-efficacy as a mediator of goal setting and performance: some human resource applications." *Journal of Managerial Psychology*, Volume 11, number 3, 33-48.

Bloom, B. et. al. (1956). *Taxonomy of educational objectives: Handbook I, Cognitive domain*. New York: McKay.

Dixon NM. (1990). "The Relationship Between Trainee Responses on Participant Reaction Forms and Posttest Scores." *Human Resource Development Quarterly*, volume 1, number 2, 129-137.

Haccoun RR, Hamtiaux T. (1994). "Optimizing knowledge test for inferring learning acquisition in single group training evaluation designs: the internal referencing strategy." *Personnel Psychology*, volume 47, number 3, 593-605.

Mann S. (1997). "Implications of the response-shift bias for management." *Journal of Management Development*, volume 16, number 5-6, 328-337.